Dual Space Preconditioning for Gradient Descent

Dual Space Preconditioning for Gradient Descent

Chris J. Maddison[1,4,*], Daniel Paulin[2,*], Yee Whye Teh[3], and Arnaud Doucet[3]

[1]University of Toronto, Toronto, Canada
[2]University of Edinburgh, Edinburgh, UK
[3]University of Oxford, Oxford, UK
[4]DeepMind, London, UK
[*]Both authors contributed equally to this work.

December 9, 2020

Abstract

The conditions of relative smoothness and relative strong convexity were recently introduced for the analysis of Bregman gradient methods for convex optimization. We introduce a generalized left-preconditioning method for gradient descent, and show that its convergence on an essentially smooth convex objective function can be guaranteed via an application of relative smoothness in the dual space. Our relative smoothness assumption is between the designed preconditioner and the convex conjugate of the objective, and it generalizes the typical Lipschitz gradient assumption. Under dual relative strong convexity, we obtain linear convergence with a generalized condition number that is invariant under horizontal translations, distinguishing it from Bregman gradient methods. Thus, in principle our method is capable of improving the conditioning of gradient descent on problems with non-Lipschitz gradient or non-strongly convex structure. We demonstrate our method on p-norm regression and exponential penalty function minimization.

1 Introduction

1.1 Setting and method

We study the minimization of a proper, closed, and essentially smooth convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$,

$$\min_{x \in \mathbb{R}^d} f(x). \tag{P}$$

For unconstrained f, i.e., $\operatorname{dom} f = \{x \in \mathbb{R}^d : f(x) < \infty\} = \mathbb{R}^d$, essential smoothness is simply differentiability. For constrained f, essential smoothness is the assumption that f is differentiable on $\operatorname{int}(\operatorname{dom} f) \neq \emptyset$ and that the norm of the gradient grows without bound, $\|\nabla f(x)\| \to \infty$, as x approaches the boundary of the domain. Thus, a global minimizer $x_{\min}$ of f, if it exists, is in $\operatorname{int}(\operatorname{dom} f)$. The method that we introduce (Algorithm 1.1) is a non-linear generalization of linear left-preconditioning for gradient descent (see, e.g., [15, sect. 9.4]), and our analysis relies on recent

Algorithm 1.1 Dual preconditioned gradient descent.

Given an essentially smooth convex $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, a Legendre convex $k : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ with $\nabla f(\operatorname{int}(\operatorname{dom} f)) \subseteq \operatorname{int}(\operatorname{dom} k)$ and $0 = \arg\min_{x^*} k(x^*)$, $x_0 \in \operatorname{int}(\operatorname{dom} f)$, and $L^* > 0$. For all $i \geq 0$,

$$x_{i+1} = x_i - \frac{1}{L^*} \nabla k(\nabla f(x_i)).$$

generalizations of the typical Lipschitz gradient assumption [7]. For the sake of exposition, we will assume in the introduction that f is twice continuously differentiable on $\operatorname{int}(\operatorname{dom} f)$, but this is not a requirement of our method.

In the analysis of first-order methods, it is standard to assume that the derivatives of f at some order are globally bounded by constants. For example, consider the gradient descent method, whose iterates satisfy

$$x_{i+1} = \underset{x \in \operatorname{dom} f}{\arg\min} \left\{ \langle \nabla f(x_i), x \rangle + \tfrac{L}{2} \|x - x_i\|^2 \right\}, \tag{1}$$

where $L > 0$ and $x_0 \in \operatorname{int}(\operatorname{dom} f)$. A classical analysis shows that the iterates of gradient descent converge linearly in i, i.e., $f(x_i) - f(x_{\min}) = \mathcal{O}(\lambda^i)$ for $\lambda = 1 - \mu/L$, when f is assumed to be $\mu > 0$ strongly convex and ∇f is assumed to be L-Lipschitz continuous (typically called "smoothness"). Taken together for twice continuously differentiable f, these conditions are equivalent to the conditions that the eigenvalues of the Hessian matrix of second-order partial derivatives $\nabla^2 f(x)$ are everywhere lower bounded by $\mu > 0$ (strong convexity) and upper bounded by $L > 0$ (smoothness),

$$\mu I \preceq \nabla^2 f(x) \preceq LI \text{ for all } x \in \operatorname{int}(\operatorname{dom} f). \tag{2}$$

Analyses of first-order methods using only non-constant bounds on the derivatives of f have recently been discovered [11, 7, 45, 42, 31]. In particular, [7] studied the following generalized gradient method that takes a designed essentially smooth, strictly convex reference function $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ with $\operatorname{int}(\operatorname{dom} f) \subseteq \operatorname{int}(\operatorname{dom} h)$. Given $x_0 \in \operatorname{int}(\operatorname{dom} f)$, this method's iterates satisfy

$$x_{i+1} = \underset{x \in \operatorname{dom} f}{\arg\min} \{ \langle \nabla f(x_i), x \rangle + L D_h(x, x_i) \} \tag{3}$$

where $L > 0$, $\langle \cdot, \cdot \rangle$ is the Euclidean inner product, and $D_h(x, y) = h(x) - h(y) - \langle \nabla h(y), x - y \rangle$ for $x, y \in \operatorname{int}(\operatorname{dom} h)$. (3) is due to [34] and falls in a family of so-called Bregman gradient methods. A standard analysis of (3) (see, e.g., [8]) makes the "absolute" assumptions that f is Lipschitz continuous and that h is strongly convex. In contrast, consider the following "relative" conditions between f and h, for $\mu \geq 0$ and $L > 0$

$$\mu \nabla^2 h(x) \preceq \nabla^2 f(x) \preceq L \nabla^2 h(x) \text{ for all } x \in \operatorname{int}(\operatorname{dom} f). \tag{4}$$

For twice continuously differentiable f, Bauschke *et al.* [7] first showed that (4) with $\mu = 0$ is a sufficient assumption to guarantee the sublinear convergence of $f(x_i) - f(x_{\min})$ in (3). Lu *et al.* [31] extended this analysis, and showed that (4) with $\mu > 0$ is sufficient for the linear convergence of $f(x_i) - f(x_{\min})$. Conditions (4) are "relative" in the sense that it is possible for (4) to hold for f and h that are both non-smooth or non-strongly convex. For example, [7] study a Poisson inverse objective whose derivatives of all orders are unbounded as $x \to 0$. They design an appropriate h,

whose Hessian is also unbounded at 0, but which satisfies (4). Analyses of first-order methods using non-constant bounds on the derivatives of f have been extended to non-convex f [12, 23] continuous convex optimization [30], composite least-squares problems [24], symmetric non-negative matrix factorization [23], and the Sinkhorn algorithm [33]. Notably, relative smoothness conditions have also been used to justify fast implementations of third-order tensor methods [36].

The method that we introduce (Algorithm 1.1) exploits an application of these relative conditions in the dual space through an essentially smooth, strictly convex dual reference function $k : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ with $\nabla f(\text{int}(\text{dom}\, f)) \subseteq \text{int}(\text{dom}\, k)$ and $0 = \arg\min_{x^*} k(x^*)$. The method is a generalization of left-preconditioned gradient descent, which we discuss in more detail in section 1.2. In section 3 we consider the conditions under which we can provide convergence rates for our method. For twice continuously differentiable f sufficient conditions that we study are the existence of $\mu^* \geq 0$, $L^* > 0$ such that

$$\mu^*[\nabla^2 k(\nabla f(x))]^{-1} \preceq \nabla^2 f(x) \preceq L^*[\nabla^2 k(\nabla f(x))]^{-1} \text{ for all } x \in \text{int}(\text{dom}\, f). \tag{5}$$

When $\mu^* = 0$, we show that $k(\nabla f(x_i)) - k(0)$ converges sub-linearly with rate $\mathcal{O}(i^{-1})$ (and thus $x_i \to x_{\min}$) along the iterates of Algorithm 1.1. When f is strictly convex and $\mu^* > 0$, we show that $f(x_i) - f(x_{\min})$ converges linearly with rate $\lambda^* = 1 - \mu^*/L^*$. As we show in section 3, assumptions (5) are relative smoothness and strong convexity assumptions in the dual space, and they are distinct from (4). In section 4, we design dual reference functions for p-norm regression (see [17, 2] and references therein) and exponential penalty functions (see, e.g., [21, 20]).

1.2 Preconditioning

In this paper, we introduce a generalization of linear left-preconditioning, which is a fundamental technique used in algorithms for solving linear systems. In this subsection, we review linear preconditioning, following closely Wathen's short introduction [46], and give an interpretation of our method and Bregman gradient methods as left- and right-preconditioning, respectively.

Consider the problem of minimizing a positive-definite quadratic, which is equivalent to finding the solution x of a linear system of d equations with d unknowns: $Ax = b$ where $b \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d \times d}$ is symmetric and positive-definite. "Preconditioning" refers to the idea of modifying this system in a way that preserves the solution, but improves the convergence of iterative methods. For example, given a positive-definite $P \in \mathbb{R}^{d \times d}$, we may consider the following systems (known as left- or right-preconditioning, respectively):

$$P^{-1}Ax = P^{-1}b \qquad \text{or} \qquad AP^{-1}y = b \text{ s.t. } x = P^{-1}y \tag{6}$$

These have the same solution as the original, and if $P^{-1}A$ or AP^{-1} approximates the identity, then iterative methods will converge faster. Indeed, for iterates of the conjugate gradients method (CG) [26], $\langle x - x_i, A(x - x_i)\rangle$ converges linearly with a rate that varies monotonically with the condition number $\kappa^A = \lambda^A_{\max}/\lambda^A_{\min}$, i.e., the ratio of the largest to the small eigenvalue of A [25, Chap. 3.1]. Smaller condition number is better, so if $\kappa^A \gg \kappa^{P^{-1}A}$, then left-preconditioned CG will converge faster. Preconditioned methods typical solve a system with P at every iteration. Thus, P should satisfy two criteria: $\kappa^{P^{-1}A}$ should be small and $Px = b$ should be easy to solve. It may seem difficult to strike this balance, but it is possible in many cases. Wathen [46] gives an example due to Strang for Toeplitz matrices that reduces the complexity of linear solves from $\mathcal{O}(d^2)$ to $\mathcal{O}(d \log d)$. More generally, preconditioners are considered essential in solvers for very large, sparse, linear systems [10, 39].

Consider now the more general problem (P) for an unconstrained f. One can show that the following are stationary conditions of Algorithm 1.1 or (3), respectively:

$$\nabla k(\nabla f(x)) = 0 \qquad \text{or} \qquad \nabla f(\nabla h^*(y)) = 0 \text{ s.t. } x = \nabla h^*(y), \tag{7}$$

where $\nabla h^*(y) = \arg\max_{x \in \mathbb{R}^d} \langle x, y \rangle - h(x)$. Clearly, (6) specializes (7) for appropriately chosen quadratic f, k, h. Thus, our method and the Bregman gradient method (3) may be seen as a generalization of left- and right-preconditioning for gradient descent, respectively. Moreover, for symmetric, positive-definite $A, P \in \mathbb{R}^{d \times d}$, the existence of $L, \mu > 0$ such that $\mu P \preceq A \preceq LP$ guarantees $\kappa^{P^{-1}A} \leq L/\mu$ and an error bound on preconditioned CG. This is generalized by the primal (4) and dual (5) relative conditions. However, in contrast to the linear case, the choice of left (dual) vs. right (primal) in the non-linear case is much more consequential and the two methods are not equivalent in general (left- and right-preconditioning for CG are equivalent [39, Chap. 9.1]). The class of f satisfying the dual conditions (5) for a fixed k is closed under horizontal translations. This is not true in general for f satisfying the primal conditions (4) for a fixed h. Thus, in general, $\mu \neq \mu^*$, $L \neq L^*$, and the global information encoded in the dual reference function k is distinct from the information encoded in the reference function h.

Non-linear preconditioning is far less studied, but has been considered in a number of works. Non-linear preconditioning methods have recently been shown to stabilize Euler discretization schemes of stochastic differential equations [27, 40]. In fact, the non-linear preconditioning of [27] is the same as the one we consider for exponential penalty functions. Finally, recent work [18, 22] developed non-linear preconditioning schemes for Newton's method applied to problems arising from the discretization of partial differential equations.

2 Convex analysis background

2.1 Essential smoothness and convex conjugates

In this section we review some basic facts of convex analysis that will be used throughout. Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper closed convex function with domain $\operatorname{dom} h = \{x : \mathbb{R}^d : h(x) < \infty\}$. To indicate $\operatorname{dom} h = \mathbb{R}^d$, we simply define $h : \mathbb{R}^d \to \mathbb{R}$ as ranging only over the reals. $\partial h(x)$ denotes the subdifferential of h at $x \in \mathbb{R}^d$. For a proper convex functions, being closed is equivalent to being lower semi-continuous (lsc). Let $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ indicate the Euclidean norm and inner product, respectively, unless otherwise specified. The convex conjugate $h^* : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ of a proper closed convex function h is given by

$$h^*(x^*) = \sup\{\langle x, x^* \rangle - h(x) : x \in \operatorname{dom} h\}. \tag{8}$$

h^* is also a proper closed convex function, and $(h^*)^* = h$ [38, Cor. 12.2.1]. For more on h^*, we refer readers to [38, 15, 13].

In this work, we study the minimization of an essentially smooth convex function f [38, Chap. 25], which can be thought of as an assumption of differentiability. For constrained f, essential smoothness comes with additional structure that prevents f from having sharp edges at the boundary of its domain. In some cases, we will consider the additional assumption that f is strictly convex on the interior of its domain.

Definition 2.1 (Essential smoothness and Legendre convexity)**.** *Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper closed convex function. h is essentially smooth if,*

1. $\operatorname{int}(\operatorname{dom} h)$ *is not empty.*

2. h *is differentiable on* $\operatorname{int}(\operatorname{dom} h)$*, with* $\lim_{i\to\infty} \|\nabla h(x_i)\| = \infty$ *whenever* $x_i \in \operatorname{int}(\operatorname{dom} h)$ *is a sequence converging to the boundary of* $\operatorname{int}(\operatorname{dom} h)$*.*

h is Legendre convex, if additionally

3. h *is strictly convex on* $\operatorname{int}(\operatorname{dom} h)$

In this work, the assumption that h is essentially smooth carries with it the implied assumption that h is proper and closed.

Essentially smooth convex functions can only be minimized in their interior.

Lemma 2.2. *If $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is an essentially smooth convex function that is minimized at $x_{\min} \in \operatorname{dom} h$, then $x_{\min} \in \operatorname{int}(\operatorname{dom} h)$.*

Proof. Suppose that $x_{\min}$ is a boundary point. Since $\operatorname{int}(\operatorname{dom} h) \neq \emptyset$, by convexity there exists a line segment connecting the boundary point $x_{\min}$ and any other interior point a. However, by [38, Lem. 26.2], we know that the directional derivative converges to $-\infty$ as we tend towards the boundary point on this line segment, hence $x_{\min}$ could not be a minimum of h. □

Legendre convex functions (essentially smooth, strictly convex functions) have even more convenient structure. One consequence of Legendre structure, which will be used in our analysis to show that k is radially unbounded, is that achieving a minimum is sufficient to imply that a Legendre convex function grows without bound.

Lemma 2.3. *Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a Legendre convex function that is minimized at $0 \in \operatorname{dom} h$. Then h is radially unbounded, i.e., if $x_i \in \mathbb{R}^d$ is a sequence such that $\|x_i\| \to \infty$, then $h(x_i) \to \infty$.*

Proof. First, by Lemma 2.2 it follows that $0 \in \operatorname{int}(\operatorname{dom} h)$. Because h is strictly convex, 0 is the unique minimum of h. Thus, we can define the sphere $\mathcal{S} = \{x \in \mathbb{R}^d : \|x\| = r\}$ for some $r > 0$ such that $\mathcal{S} \in \operatorname{int}(\operatorname{dom} h)$. By continuity of h in the interior of its domain, and the uniqueness of the minimum at zero, we have $\inf_{x\in\mathcal{S}} h(x) > h(0)$. Now, assume without loss of generality that $\|x_i\| > r$. By strict convexity of Legendre functions, property 3, we have

$$h(0) + \frac{\|x_i\|}{r}\left(h\left(\frac{r x_i}{\|x_i\|}\right) - h(0)\right) < h(0) + (h(x_i) - h(0)) \tag{9}$$

and thus

$$h(x_i) > h(0) + \frac{\|x_i\|}{r}\left(\inf_{x\in\mathcal{S}} h(x) - h(0)\right). \tag{10}$$

Our result follows by taking $i \to \infty$. □

A second key consequence of Legendre structure is that the gradient map ∇h is invertible and given by $(\nabla h)^{-1} = \nabla h^*$, which also gives a characterization of the inverse of $\nabla^2 h(x)$. We summarize both of these properties in Lemma 2.4.

Lemma 2.4. *Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be Legendre convex. Then, h^* is Legendre convex, the map ∇h is one-to-one and onto from the open set $\operatorname{int}(\operatorname{dom} h)$ onto the open set $\operatorname{int}(\operatorname{dom} h^*)$, continuous in both directions, and for all $x \in \operatorname{int}(\operatorname{dom} h)$*

$$\nabla h^*(\nabla h(x)) = x. \tag{11}$$

If h is C^2 on an open set containing x and $\det \nabla^2 h(x) \neq 0$, then

$$\nabla^2 h^*(\nabla h(x)) \nabla^2 h(x) = \nabla^2 h(x) \nabla^2 h^*(\nabla h(x)) = I. \tag{12}$$

Proof. For the first part see Rockafellar [38, Thm. 26.5]. For (12), note that, by the inverse function theorem, ∇h^* is continuously differentiable at $\nabla h(x)$ under the assumption that ∇h is continuously differentiable on an open set containing x. The remainder follows by the chain rule applied to (11). □

2.2 Relative smoothness and relative strong convexity

Analyses of first-order methods for differentiable optimization typically require that ∇f is Lipschitz continuous (smooth). Recent generalizations of smoothness (and strong convexity) [7, 31] can be used to guarantee convergence of first-order methods beyond the Lipschitz ∇f case. We will use these in our analysis of dual preconditioning. Following [7], we define these relative conditions in terms of zeroth-order properties.

Definition 2.5 (Relative smoothness and strong convexity)**.** *Let $h, g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be proper closed convex functions, $Q \subseteq \operatorname{dom} h \cap \operatorname{dom} g$ be a convex set, and $L, \mu \geq 0$. Define $d_L, d_\mu : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ for $x \in Q$ by*

$$d_L(x) = Lg(x) - h(x) \qquad d_\mu(x) = h(x) - \mu g(x) \tag{13}$$

and for $x \notin Q$ by $d_L(x) = d_\mu(x) = \infty$. h is L-smooth relative to g on Q, if d_L convex. h is μ-strongly convex relative to g on Q, if d_μ is convex.

The special cases with $g(x) = \|x\|_2^2 / 2$ are exactly the classical conditions of strong convexity and smoothness. We now provide first- and second-order characterizations.

2.3 First-order characterizations for relative conditions

The first-order characterizations of relative smoothness and strong convexity are given in terms of the Bregman divergence [16, 6], which for essentially smooth convex $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and $x \in \operatorname{dom} h, y \in \operatorname{int}(\operatorname{dom} h)$ is given by $h(x) - h(y) - \langle \nabla h(y), x - y \rangle$. Unfortunately, in our analysis, we will require smoothness relative to f^*, which can fail to be differentiable when f is essentially smooth. Thus, we will make use of a generalization of the Bregman divergence, which we define via the *one-sided directional derivative* of $h : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, \infty\}$ with respect to $y \in \mathbb{R}^d$ at a point x where h is finite:

$$h'(x; y) = \lim_{\epsilon \downarrow 0} \frac{h(x + \epsilon y) - h(x)}{\epsilon}. \tag{14}$$

This may take values in $\{\infty, -\infty\}$. The advantage of one-sided direction derivatives is that they always exist at $x \in \operatorname{dom} h$ for proper convex h [38, Thm. 23.1]. We are now prepared to define a novel generalization of the Bregman divergence.

Definition 2.6 (Generalized Bregman divergences)**.** *Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, \infty\}$ be a function. Let $x, y \in \mathbb{R}^d$ be points at which h is finite and $h'(y; x - y)$ exists. Define the generalized Bregman divergence,*

$$D_h(x, y) = h(x) - h(y) - h'(y; x - y). \tag{15}$$

If h is proper, closed, and convex, then $D_h(x, y)$ is defined for all $x, y \in \operatorname{dom} h$ [38, Thm. 23.1] and is finite for $y \in \operatorname{dom} \partial h = \{x \in \mathbb{R}^d : \partial h(x) \neq \emptyset\}$ [38, Thm. 23.2].

Clearly, $D_h(x, y)$ coincides with the standard Bregman divergence if h is differentiable at y. The advantage of the generalization is that it allows us to define the relative conditions in terms of first-order properties without the assumption of differentiability.

Proposition 2.7 (First-order characterizations of relative conditions)**.** *Let $h, g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be proper closed convex functions, $Q \subseteq \operatorname{dom} h \cap \operatorname{dom} g$ be a convex set, and $L, \mu \geq 0$. The following are equivalent*

1. *h is L-smooth relative to g on Q.*
2. *For all $x, y \in Q$, $D_h(x, y) \leq L D_g(x, y)$.*

The following are equivalent

3. *h is μ-strongly convex relative to g on Q.*
4. *For all $x, y \in Q$, $\mu D_g(x, y) \leq D_h(x, y)$.*

To prove these equivalences, we will use two lemmas, which extend the first-order characterization of one-dimensional convexity to the non-differentiable case.

Lemma 2.8 (A variant of the mean value theorem)**.** *Let $r \colon [0, 1] \to \mathbb{R}$ be a continuous function. Define $r'_+(z) = \lim_{\epsilon \downarrow 0} (r(z + \epsilon) - r(z))/\epsilon$. Assuming that $r'_+(z)$ exists for $z \in [0, 1)$, if $s, t \in [0, 1]$ and $s < t$, then there is exists $z \in [s, t)$ such that $r'_+(z) \geq (r(t) - r(s))/(t - s)$.*

Proof. We can add a linear function to r without changing the difference between the two sides on the inequality, so without loss of generality we may assume that $r(s) = r(t) = 0$. Then, we want to prove that $r'_+(z) \geq 0$ for some $z \in [s, t]$. Since r is continuous, it has a minimum in $[s, t]$, so there is a $z \in [s, t]$ such that $r(u) \geq r(z)$ for every $u \in [s, t]$. If $z = t$, then $r(z) = r(s)$, so we could instead take $z = s$. Thus we may assume that $z \in [s, t)$. Then $r'_+(z) = \lim_{u \downarrow z,\, u \in (z,t)} \frac{r(u) - r(z)}{u - z} \geq 0$, because $r(u) \geq r(z)$, proving the claim. □

Lemma 2.9 (Characterization of one-dimensional convexity)**.** *Let $r \colon [0, 1] \to \mathbb{R}$ be a continuous function. r is convex on $[0, 1]$ if and only if $r'_+(z)$ (defined in Lemma 2.8) exists for all $z \in (0, 1)$, and for all $s, t \in (0, 1)$ such that $s < t$,*

$$r(t) \geq r(s) + r'_+(s)(t - s). \tag{16}$$

Proof. Suppose that r is not convex on $[0, 1]$. Then by continuity, it is also not convex on $(0, 1)$. After adding a linear function, we can arrange that $0 = r(s) = r(t) < r(z)$ for some $0 < s < z < t < 1$. Since r is continuous, r achieves its maximum restricted to the interval $[s, t]$, so we could choose $z \in (s, t)$ such that $r(z) = \max_{u \in [s,t]} r(u) > 0$. Since $r(z) > 0$ and r is continuous, there is a $u \in (s, z)$ such that $r(v) > 0$ for every $v \in [u, z]$. By Lemma 2.8, there is a $v \in [u, z)$ such that

$r'_+(v) \geq \frac{r(z)-r(u)}{z-u} \geq 0$, and so $0 = r(t) \geq r(v) + r'_+(v)(t-v) \geq r(v) > 0$. This contradiction proves that r is indeed convex.

Now suppose that r is convex and continuous on $[0,1]$. By [38, Thm. 23.1] the difference quotient $\frac{r(z+\epsilon)-r(z)}{\epsilon}$ is a non-decreasing function of ϵ for $\epsilon > 0$ and $z \in [0,1)$, and limit $r'_+(z)$ exists. Using the fact that the difference quotient is non-decreasing in ϵ it follows that $r(t) \geq r(s) + r'_+(s)(t-s)$ for every $s, t \in [0,1]$, $s < t$. □

We are now prepared to provide the proof of equivalence between the zeroth- and first-order definitions of the relative conditions.

Proof of Proposition 2.7. We only show the equivalence of 1. and 2., the proof of the equivalence of 3. and 4. is similar. First, suppose that 1. holds, i.e., d_L is convex. Let $x, y \in Q$. Then for $x_t = y + t(x-y)$, we have (after dividing by t and rearranging)

$$h(x) - h(y) - \frac{h(x_t) - h(y)}{t} \leq Lg(x) - Lg(y) - L\frac{g(x_t) - g(y)}{t} \tag{17}$$

Taking the limit $t \downarrow 0$ gives us that $D_h(x,y) \leq LD_g(x,y)$, with the existence of the limits following from [38, Thm. 23.1].

For the other direction, suppose that $D_h(x,y) \leq LD_g(x,y)$ for every $x, y \in Q$. Let $x, y \in Q$, and $x_t = y + t(x-y)$. Then for any $0 < s < t < 1$, it is easy to check that both $h'(x_t; x_s - x_t)$ and $g'(x_t; x_s - x_t)$ are finite (if one of the directional derivatives is non-finite, then this would contradict the convexity or finiteness of these functions over Q). Thus $D_h(x_s, x_t)$ and $D_g(x_s, x_t)$ are finite and satisfy $D_h(x_s, x_t) \leq LD_g(x_s, x_t)$. Thus, it follows that for any $0 < s < t < 1$,

$$D_{d_L}(x_s, x_t) = LD_g(x_s, x_t) - D_h(x_s, x_t) \geq 0. \tag{18}$$

Let $r(t) = d_L(x_t)$ for $t \in [0,1]$ be the restriction of d_L on the line segment between x and y, then (18) implies that the condition (16) holds. r is a continuous function by [38, Thm. 10.2]. Thus, by Lemma 2.9, r is a convex function on $[0,1]$. This holds for all $x, y \in Q$, thus d_L is convex.

□

2.4 Second-order characterizations of relative conditions

Verifying relative smoothness or strong convexity is typically done via second-order conditions. Just as Lipschitz continuity of ∇h can be characterized by a bound on $\nabla^2 h$, the relative conditions can be characterized by the second derivatives of h and g [7, 31]. Proposition 2.10 allows $\nabla^2 h, \nabla^2 g$ to be undefined at a point, a slight generalization of the standard result that is useful in our analysis when $\nabla^2 f$ is undefined at $x_{\min}$.

Proposition 2.10 (Second-order characterizations of relative conditions)**.** *Let $h, g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be proper closed convex functions that are differentiable on the interior of their domains. Let $Q \subseteq \operatorname{int}(\operatorname{dom} g) \cap \operatorname{int}(\operatorname{dom} h)$ be an open convex set, $z \in Q$, and $L, \mu \geq 0$. If h, g are C^2 on $Q \setminus \{z\}$, then*

1. *h is L-smooth relative to g on Q if and only if,*

$$\nabla^2 h(x) \preceq L\nabla^2 g(x) \qquad \forall x \in Q \setminus \{z\}.$$

2. h is μ-strongly convex relative to g on Q if and only if,

$$\mu\nabla^2 g(x) \preceq \nabla^2 h(x) \qquad \forall x \in Q \setminus \{z\}.$$

Proof. Again, we prove the relative smoothness equivalence, and relative strong convexity follows similarly. For relative smoothness, ($\Rightarrow$) follows from part one of [35, Thm. 2.1.4] applied to d_L at $x \in Q$. For ($\Leftarrow$), it is sufficient to prove convexity of the restriction of d_L to an open line segment with endpoints $x, y \in Q$. Let $x_t = y + t(x-y)$ and $r(t) = d_L(x_t)$ for $t \in (0,1)$. Let $a \in (0,1)$ be such that $x_a = z$, if it exists, or some arbitrary $a \in (0,1)$, otherwise. d_L is continuously differentiable at all $x \in Q$ by [38, Thm 25.5]. Thus $r'(t) = \langle \nabla d_L(x_t), x - y\rangle$ is a continuous and finite function of $t \in (0,1)$. If r' is non-decreasing, then the argument of [38, Thm. 4.4] gives us our result. Thus, with a slight abuse of notation,

$$r'(t) = r'(a) + r'(t) - r'(a) = r'(a) + \lim_{s\to a} \int_s^t \left\langle x - y, \nabla^2 d_L(x_t)(x-y)\right\rangle.$$

The limit is actually a one-sided limit, depending on $t \le a$ or $t > a$. Either way, $\nabla^2 d_L(x_t) = L\nabla^2 g(x_t) - \nabla^2 h(x_t)$ is positive semi-definite, so r' is non-decreasing. □

3 Analysis of the dual preconditioned scheme

3.1 Motivation and assumptions

Relative smoothness of f with respect to a reference function h is the key assumption under which [7, 42, 31] analyzed the convergence of Bregman gradient methods. We now build towards an analysis of dual space preconditioned gradient descent method (Algorithm (1.1)) using the assumption that k is smooth relative to f^*. As shorthand to distinguish these two assumptions, we use the terms *primal relative smoothness* to refer to the condition that f is L-smooth relative to h and *dual relative smoothness* to refer to the condition that k is L^*-smooth relative to f^*. To motivate our assumption, consider the following idealizations.

Consider the Bregman gradient method update (3), which can be rewritten as,

$$x_{i+1} = \arg\min_{x\in \operatorname{dom} f} \left\{\langle \nabla f(x_i) - L\nabla h(x_i), x\rangle + Lh(x)\right\}. \tag{19}$$

In this form, it is clear that, if $h = f$ and $L = 1$, then the iteration would converge in a single step to the minimizer of $h = f$. This is an idealization, because a single iteration would be as expensive to compute as the original problem. The spirit behind primal relative smoothness is that the condition $h = f$ can be relaxed to admit h for which the update (19) is efficiently solvable and the iterates still converge.

Now, consider the case that f is Legendre convex with a minimum at $x_{\min}$, and let $f_c^*(x^*) = f^*(x^*) - \langle x^*, x_{\min}\rangle$ for $x^* \in \mathbb{R}^d$. Notice that $\nabla f_c^*(\nabla f(x)) = x - x_{\min}$ by Lemma 2.4 and that Algorithm 1.1 with $k = f_c^*$ and $L_i = 1$ would converge in a single step to the minimizer $x_{\min}$ of f. Thus, in analogy to the relative smoothness analysis of [7] in the primal space, the spirit behind our analysis under dual relative smoothness is that the requirement $k = f_c^*$ can be relaxed while maintaining the convergence of Algorithm 1.1. In particular, sufficient assumptions on k are that it is minimized at 0 and smooth relative to f^*.

More precisely, our analysis of Algorithm 1.1 uses following assumptions.

Assumption 3.1. *1. $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is convex and essentially smooth.*

2. $k : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is Legendre convex and uniquely minimized at 0.

3. $\nabla f(\operatorname{int}(\operatorname{dom} f)) \subseteq \operatorname{int}(\operatorname{dom} k)$ and for all $x, y \in \operatorname{int}(\operatorname{dom} f)$,

$$D_k(\nabla f(y), \nabla f(x)) \leq L^* D_f(x, y)$$

As we show in the following sections, Assumption 3.1.3 is a necessary condition of the relative smoothness of k with respect to f^*. Assumption 3.1.3 is the assumption that requires the most effort to verify, since the convexity of $L^* f^* - k$ will typically be difficult to check. For this reason, we also provide second-order sufficient conditions expressed (mostly) in terms of conditions on f and k.

3.2 Dual relative conditions for Legendre convex objectives

When f is essentially smooth and strictly convex (Legendre), we are able to provide clean characterizations of the dual relative conditions. In particular, Assumption 3.1.3 is necessary and sufficient for the smoothness of k relative to f^* on $\operatorname{int}(\operatorname{dom} f^*)$. We begin by linking D_f and D_{f^*} in what is a well-known identity for Legendre convex f.

Lemma 3.2. *If $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is an essentially smooth convex function, then*

$$D_{f^*}(\nabla f(y), \nabla f(x)) \leq D_f(x, y) \tag{20}$$

for all $x, y \in \operatorname{int}(\operatorname{dom} f)$. If f is Legendre convex, then this is an equality.

Proof. Note, by [38, Cor. 26.4.1], we have $\nabla f(\operatorname{int}(\operatorname{dom} f)) = \operatorname{dom} \partial f^*$. Thus, $D_{f^*}(\nabla f(y), \nabla f(x))$ is finite for any $x, y \in \operatorname{int}(\operatorname{dom} f)$. Note $x \in \partial f^*(\nabla f(x))$. Now,

$$\begin{aligned} D_{f^*}(\nabla f(y), \nabla f(x)) &= f^*(\nabla f(y)) - f^*(\nabla f(x)) - (f^*)'(\nabla f(x); \nabla f(y) - \nabla f(x)) \\ &\overset{(a)}{\leq} f^*(\nabla f(y)) - f^*(\nabla f(x)) - \langle x, \nabla f(y) - \nabla f(x) \rangle \\ &\overset{(b)}{=} -f(y) + \langle \nabla f(y), y \rangle + f(x) - \langle \nabla f(x), x \rangle - \langle x, \nabla f(y) - \nabla f(x) \rangle \\ &= f(x) - f(y) + \langle \nabla f(y), y - x \rangle = D_f(x, y) \end{aligned}$$

where (a) follows from [38, Thm. 23.2], (b) follows from [38, Thm. 26.4]. If f is Legendre convex, then by Lemma 2.4, f^* is Legendre, f^* is differentiable on $\operatorname{int}(\operatorname{dom} f^*) = \nabla f(\operatorname{int}(\operatorname{dom} f))$, and (a) is an equality [38, Thm. 23.4] . □

We can now provide first-order characterizations of the dual relative conditions.

Proposition 3.3 (First-order characterization of dual relative conditions, Legendre convex case)**.** *Let $f, k : \mathbb{R}^d \to \{\mathbb{R}, \infty\}$ be Legendre convex functions. The following are equivalent.*

1. k is L^*-smooth relative to f^* on $\operatorname{int}(\operatorname{dom} f^*)$.

2. $\nabla f(\operatorname{int}(\operatorname{dom} f)) \subseteq \operatorname{int}(\operatorname{dom} k)$, and for all $x, y \in \operatorname{int}(\operatorname{dom} f)$,

$$D_k(\nabla f(y), \nabla f(x)) \leq L^* D_f(x, y).$$

The following are equivalent.

3. *k is μ^*-strongly convex relative to f^* on* $\operatorname{int}(\operatorname{dom} f^*)$.

4. $\nabla f(\operatorname{int}(\operatorname{dom} f)) \subseteq \operatorname{int}(\operatorname{dom} k)$, *and for all* $x, y \in \operatorname{int}(\operatorname{dom} f)$,

$$\mu^* D_f(x, y) \leq D_k(\nabla f(y), \nabla f(x)).$$

Proof. We prove the relative smoothness results, and the relative strong convexity ones follow similarly. First, notice that that $\nabla f(\operatorname{int}(\operatorname{dom} f)) = \operatorname{int}(\operatorname{dom} f^*)$. by Lemma 2.4. For $(1 \Rightarrow 2)$, by the definition of relative smoothness, we have $\operatorname{int}(\operatorname{dom} f^*) \subseteq \operatorname{dom} k$, and since this is an open set, we necessarily have $\operatorname{int}(\operatorname{dom} f^*) \in \operatorname{int}(\operatorname{dom} k)$. By Proposition 2.7 we have that for all $x^*, y^* \in \operatorname{int}(\operatorname{dom} f^*)$,

$$D_k(y^*, x^*) \leq L^* D_{f^*}(y^*, x^*). \tag{21}$$

By Lemmas 2.4 and 3.2, this implies

$$D_k(\nabla f(y), \nabla f(x)) \leq L^* D_{f^*}(\nabla f(y), \nabla f(x)) = L^* D_f(x, y), \tag{22}$$

for all $x, y \in \operatorname{int}(\operatorname{dom} f)$. For $(2 \Rightarrow 1)$, by Lemma 3.2, we have for all $x, y \in \operatorname{int}(\operatorname{dom} f)$,

$$D_f(x, y) = D_{f^*}(\nabla f(y), \nabla f(x)). \tag{23}$$

Using this, Proposition 2.7 implies that k is L^*-smooth relative to f^* on $\operatorname{int}(\operatorname{dom} f^*)$. □

If f is Legendre convex, then the dual relative conditions have a natural second-order characterization, which reveals the structure of the difference between them and primal relative conditions. Again, typically it is easiest to prove dual relative smoothness (or strong convexity) via these second-order conditions.

Proposition 3.4 (Second-order characterizations of dual relative conditions, Legendre convex case)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be Legendre convex, minimized at $x_{\min}$, and C^2 on* $\operatorname{int}(\operatorname{dom} f) \setminus \{x_{\min}\}$ *such that* $\det \nabla^2 f(x) \neq 0$ *at* $x \in \operatorname{int}(\operatorname{dom} f) \setminus \{x_{\min}\}$. *Let $k : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be Legendre convex, C^2 on* $\operatorname{int}(\operatorname{dom} f^*) \setminus \{0\}$ *such that* $\det \nabla^2 k(x^*) \neq 0$ *at* $x^* \in \operatorname{int}(\operatorname{dom} f^*) \setminus \{0\}$. *Let $L, \mu \geq 0$.*

1. *k is L^*-smooth relative to f^* on* $\operatorname{int}(\operatorname{dom} f^*)$ *if and only if,*

$$\nabla^2 f(x) \preceq L^* [\nabla^2 k(\nabla f(x))]^{-1} \qquad \forall x \in \operatorname{int}(\operatorname{dom} f) \setminus \{x_{\min}\}.$$

2. *k is μ^*-strongly convex relative to f^* on* $\operatorname{int}(\operatorname{dom} f^*)$ *if and only if,*

$$\mu^* [\nabla^2 k(\nabla f(x))]^{-1} \preceq \nabla^2 f(x) \qquad \forall x \in \operatorname{int}(\operatorname{dom} f) \setminus \{x_{\min}\}.$$

Remark 3.5. It is well-known that the primal and dual relative conditions are equivalent in the case of $\nabla^2 h(x) = I = \nabla^2 k(x^*)$ (see, e.g., [48, 41, 28]). In particular, if f is μ-strongly convex and L-smooth on $\operatorname{int}(\operatorname{dom} f)$, then its convex conjugate f^* is $(1/L)$-strongly convex and $(1/\mu)$-smooth on $\operatorname{int}(\operatorname{dom} f^*)$. In fact, for twice continuously differentiable f, the equivalence is a simple consequence of Propositions 2.10 and 3.4. However, this equivalence is not true in general.

Given a Legendre convex $g : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ define the following sets of functions

$$\mathscr{F}_g = \{\text{Legendre convex } f : f \text{ is smooth and strongly convex relative to } g\}, \tag{24}$$
$$\mathscr{F}_g^* = \{\text{Legendre convex } f : g \text{ is smooth and strongly convex relative to } f^*\}. \tag{25}$$

Let $k(x^*) = |x^*|^q/q$ for $x^* \in \mathbb{R}$ and $1 < q < 2$. A simple argument by contradiction shows that $\mathscr{F}_k^* \nsubseteq \mathscr{F}_h$ for all twice continuously differentiable $h : \mathbb{R} \to \mathbb{R}$, implying that the primal and dual relative conditions are not equivalent in general. Consider

$$f_b(x) = |x - b|^p/p, \tag{26}$$

for $p = \frac{q}{q-1}$ and $x \in \mathbb{R}$. First $f_b \in \mathcal{F}_k^*$ for all b, which follows from $[k''(f_b'(x))]^{-1} = (p-1)|x - b|^{p-2} = f_b''(x)$ and Proposition 3.4. On the other hand, suppose there is some twice continuously differentiable $h : \mathbb{R} \to \mathbb{R}$ such that $f_b \in \mathscr{F}_h$ for all b. Then there exists $\mu > 0$ such that $\mu h''(b) \leq f_b''(b) = 0$ for all b. This implies that $h''(x) \equiv 0$ and thus $h(x) \equiv 0$. However, this leads to a contradiction, because smoothness is violated: $f_b''(b + \epsilon) > 0 = Lh''(x)$ for any $L, \epsilon > 0$.

Proof of Proposition 3.4. We prove the relative smoothness result, and the relative strong convexity one follows similarly. By Lemma 2.4, if ∇f is continuously differentiable for $x \in \operatorname{int}(\operatorname{dom} f) \setminus \{x_{\min}\}$, then ∇f^* is continuously differentiable for $x^* \in \operatorname{int}(\operatorname{dom} f^*) \setminus \{0\}$ by the inverse function theorem. Thus, by Proposition 2.10 dual relative smoothness is equivalent to: for all $x^* \in \operatorname{int}(\operatorname{dom} f^*) \setminus \{0\}$,

$$\nabla^2 k(x^*) \preceq L^* \nabla^2 f^*(x^*). \tag{27}$$

By Lemma 2.4, (27) is equivalent to for all $x \in \operatorname{int}(\operatorname{dom} f) \setminus \{x_{\min}\}$,

$$\nabla^2 k(\nabla f(x)) \preceq L^* [\nabla^2 f(x)]^{-1}. \tag{28}$$

Since $A^{-1} \preceq B^{-1}$ is equivalent to $B \preceq A$ for positive definite matrices, we are done. □

A major difference between the primal and dual relative conditions is the fact that dual relative conditions are invariant under horizontal translations of f. To see why, let k be L^*-smooth relative to f^* on a convex set Q. Define $g(x) = f(x - z)$ for $z \in \mathbb{R}^d$. Then, by [38, Thm. 12.3], $g^*(x^*) = f^*(x^*) + \langle z, x^* \rangle$. Bregman divergences of functions that differ only in affine terms are identical (see [6] for the differentiable case), so we have for all $x^*, y^* \in Q$, $D_k(x^*, y^*) \leq L^* D_{f^*}(x^*, y^*) = L^* D_{g^*}(x^*, y^*)$. Thus k is L^*-smooth relative to g^* on Q. Invariance under horizontal translation is clearly easy to violate in the case of primal relative smoothness.

Even if h is allowed to translate with f, the primal and dual relative conditions can lead to distinct conditioning. Given a positive definite $A \succ 0$, let

$$f(x) = \|Ax - b\|^p/p, \qquad h(x) = \left\|x - A^{-1}b\right\|^p/p, \qquad k(x^*) = \|x^*\|^q/q, \tag{29}$$

for $1/p + 1/q = 1$ and $p > 2$. It can be shown that f satisfies both the dual (with respect to k) and primal (with respect to h) relative conditions. Nonetheless, the condition numbers are distinct. A simple calculation reveals that in this case

$$\frac{L}{\mu} = p^2 \left(\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}\right)^p \qquad \text{vs.} \qquad \frac{L^*}{\mu^*} = (p-1)^2 \left(\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}\right)^{4-q}, \tag{30}$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are the smallest and largest singular values of A, respectively. Thus, the primal condition number is larger than the dual number (since $4 - q = 3 - (p-1)^{-1} < p$ when $p > 2$). Similarly, the example $f(x) = \|Ax - b\|_4^4/4 + \|Cx - d\|_2^2/2$ of [31, p. 339] can be shown to have better conditioning under the dual preconditioned method than under the Bregman gradient method.

3.3 Dual relative conditions for essentially smooth objectives

We now show that the smoothness of k relative to f^* on $\operatorname{dom} f^*$ is a sufficient condition for Assumption 3.1.3. We also provide a sufficient, second-order condition.

Proposition 3.6. *Let $f, k : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be essentially smooth convex functions. If k is L^*-smooth relative to f^* on $\operatorname{dom} f^*$, then $\nabla f(\operatorname{int}(\operatorname{dom} f)) \subseteq \operatorname{int}(\operatorname{dom} k)$ and for all $x, y \in \operatorname{int}(\operatorname{dom} f)$,*

$$D_k(\nabla f(y), \nabla f(x)) \leq L^* D_f(x, y). \tag{31}$$

Proof. Note that by [38, Cor. 26.4.1], $\nabla f(\operatorname{int}(\operatorname{dom} f)) = \operatorname{dom} \partial f^* \subseteq \operatorname{dom} f^*$. We have $\operatorname{dom} f^* \subseteq \operatorname{dom} k$ by the definition of relative smoothness. By Proposition 2.7 and Lemma 3.2, for every $x, y \in \operatorname{int}(\operatorname{dom} f)$ we have

$$D_k(\nabla f(y), \nabla f(x)) \leq L^* D_{f^*}(\nabla f(y), \nabla f(x)) \leq L^* D_f(x, y) \tag{32}$$

Now, we are going to show that $\nabla f(\operatorname{int}(\operatorname{dom} f)) \subseteq \operatorname{int}(\operatorname{dom} k)$. We argue by contradiction. Suppose there is a $x^* \in \nabla f(\operatorname{int}(\operatorname{dom} f))$ such that $x^* \notin \operatorname{int}(\operatorname{dom} k)$. So x^* has to be in $\operatorname{dom} k \setminus \operatorname{int}(\operatorname{dom} k)$, i.e., on the boundary of $\operatorname{int}(\operatorname{dom} k)$. Using the essential smoothness of k, by [38, Lemma 26.2] it follows that for any $y^* \in \operatorname{int}(\operatorname{dom} k)$, we have

$$k'(x^* + \lambda(y^* - x^*); y^* - x^*) \downarrow -\infty \text{ as } \lambda \downarrow 0. \tag{33}$$

We fix an arbitrary $y^* \in \operatorname{int}(\operatorname{dom} k)$, and define the function $h : [0, 1] \to \mathbb{R}$ as $h(\lambda) = L^* f^*(x^* + \lambda(y^* - x^*)) - k(x^* + \lambda(y^* - x^*))$. Then relative smoothness implies that h is a finite, continuous convex function on $[0, 1]$. However, for such a function we must have $\limsup_{\lambda \downarrow 0} h'_+(\lambda) < \infty$, since otherwise it could not be finite on $[0, 1]$ by Lemma 2.9. By combining this with (33), it follows that

$$f^{*\prime}(x^* + \lambda(y^* - x^*); y^* - x^*) \to \infty \text{ as } \lambda \downarrow 0.$$

Let $r : [0, 1] \to \mathbb{R}$ be $r(\lambda) = f^*(x^* + \lambda(y^* - x^*))$, then this implies that $r'_+(\lambda) \to \infty$ as $\lambda \downarrow 0$, and by (16) of Lemma 2.9, this contradicts the assumption that f^* is finite in $\operatorname{dom} f^*$. Hence we must have $x^* \in \operatorname{int}(\operatorname{dom} k)$, and $\nabla f(\operatorname{int}(\operatorname{dom} f)) \subseteq \operatorname{int}(\operatorname{dom} k)$. □

The next proposition gives a second-order sufficient condition for Assumption 3.1.3.

Proposition 3.7. *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be essentially smooth, and C^2 on $\operatorname{int}(\operatorname{dom} f)$. Let $k : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be Legendre convex, and C^2 on $\operatorname{int}(\operatorname{dom} k)$. If $\nabla f(\operatorname{int}(\operatorname{dom} f)) \subseteq \operatorname{int}(\operatorname{dom} k)$, $\det \nabla^2 k(x^*) \neq 0$ for all $x^* \in \operatorname{int}(\operatorname{dom} k)$, and*

$$\nabla^2 f(x) \preceq L^* [\nabla^2 k(\nabla f(x))]^{-1} \qquad \forall x \in \operatorname{int}(\operatorname{dom} f), \tag{34}$$

then $D_k(\nabla f(y), \nabla f(x)) \leq L^* D_f(x, y)$ for every $x, y \in \operatorname{int}(\operatorname{dom} f)$.

Proof. Let $x, y \in \operatorname{int}(\operatorname{dom} f)$, and let $W \subseteq \mathbb{R}^d$ be a bounded open neighborhood of the segment $I = [\nabla f(x), \nabla f(y)] = \{t\nabla f(x) + (1-t)\nabla f(y) : 0 \leq t \leq 1\}$ such that $\operatorname{cl}(W) \subseteq \operatorname{int}(\operatorname{dom} k)$. Let $\delta > 0$. Let $\epsilon > 0$, and let $f_\epsilon : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, $f_\epsilon(z) = f(z) + \epsilon\sqrt{1 + \|z\|^2}$. Then f_ϵ is a Legendre convex function, and $\operatorname{dom}(f_\epsilon) = \operatorname{dom}(f)$. We have $\nabla f_\epsilon(z) = \nabla f(z) + \epsilon(1 + \|z\|^2)^{-\frac{1}{2}} z$ and $\nabla^2 f_\epsilon(z) = \nabla^2 f(z) + \epsilon(1 + \|z\|^2)^{-\frac{3}{2}}((1 + \|z\|^2) I_d - z^T z) \succ 0$ for every $z \in \operatorname{int}(\operatorname{dom} f)$. So $\|\nabla f_\epsilon(z) - \nabla f(z)\| \leq \epsilon$ and $0 \preceq \nabla^2 f_\epsilon(z) - \nabla^2 f(z) \preceq \epsilon I_d$ for every $z \in \operatorname{int}(\operatorname{dom} f)$. Let I_ϵ denote the segment $[\nabla f_\epsilon(x), \nabla f_\epsilon(y)] \subseteq \mathbb{R}^d$. Choose ϵ small enough so that

1. the 2ϵ-neighborhood of I_ϵ is in W (so $\operatorname{dist}(I_\epsilon, \mathbb{R}^d \setminus W) \geq 2\epsilon$).
2. $\epsilon I_d \preceq \frac{\delta}{2}[\nabla^2 k(w)]^{-1}$ for every $w \in W$.
3. $\forall w_1, w_2 \in W$ s.t. $\|w_1 - w_2\| \leq \epsilon$, $[\nabla^2 k(w_1)]^{-1} \preceq (1 + \frac{\delta}{2L^*})[\nabla^2 k(w_2)]^{-1}$ (uniform continuity of $(\nabla^2 k)^{-1}$ on compact set $\operatorname{cl}(W)$ by Heine-Cantor theorem).

We will show that $(L^* + \delta) f_\epsilon^* - k$ is convex when restricted to the segment I_ϵ. Let $w \in I_\epsilon$ and $z = (\nabla f_\epsilon)^{-1}(w) \in \operatorname{int}(\operatorname{dom} f)$. Then $\nabla f_\epsilon(z) = w$, and since $\|\nabla f(z) - w\| \leq \epsilon$, we get $\nabla f(z) \in W$. We have $\nabla^2((L^* + \delta) f_\epsilon^* - k)(w) = (L^* + \delta)[\nabla^2 f_\epsilon(z)]^{-1} - \nabla^2 k(\nabla f_\epsilon(z))$, and we would like to show that this is $\succeq 0$. So we want to show $\nabla^2 f_\epsilon(z) \preceq (L^* + \delta)[\nabla^2 k(\nabla f_\epsilon(z))]^{-1}$. This follows from $\nabla^2 f_\epsilon(z) \preceq \nabla^2 f(z) + \epsilon I_d$ and $\epsilon I_d \preceq \frac{\delta}{2}[\nabla^2 k(\nabla f_\epsilon(z))]^{-1}$ and $\nabla^2 f(z) \preceq L^*[\nabla^2 k(\nabla f(z))]^{-1} \preceq (L^* + \frac{\delta}{2})[\nabla^2 k(\nabla f_\epsilon(z))]^{-1}$. So for small enough ϵ's $(L^* + \delta) f_\epsilon^* - k$ is indeed convex when restricted to I_ϵ. Then $D_k(\nabla f_\epsilon(y), \nabla f_\epsilon(x)) \leq (L^* + \delta) D_{f_\epsilon^*}(\nabla f_\epsilon(y), \nabla f_\epsilon(x)) = (L^* + \delta) D_{f_\epsilon}(x, y)$, using the convexity of $(L^* + \delta) f_\epsilon^* - k$ on I_ϵ combined with the same limiting argument as in (17), and Lemma 3.2. Taking $\epsilon \downarrow 0$, and then $\delta \downarrow 0$ we get $D_k(\nabla f(y), \nabla f(x)) \leq L^* D_f(x, y)$. □

3.4 Convergence rates for dual space preconditioned gradient descent

In this section we show that Assumption 3.1 is sufficient to provide convergence rates for Algorithm 1.1 on essentially smooth convex f. We find that $k(\nabla f(x_i))) - k(0)$ converges with rate $\mathcal{O}(i^{-1})$. Under an additional dual relative strong convexity condition, we find that $f(x_i) - f(x_{\min})$ converges with rate $\mathcal{O}((1 - \mu^*/L^*)^i)$. We begin with the following descent lemma.

Lemma 3.8 (Descent lemma)**.** *Let $f, k : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ satisfy Assumption 3.1. If $x_0 \in \operatorname{int}(\operatorname{dom} f)$, then for all $i > 0$, the iterates x_i of Algorithm 1.1 are such that $x_i \in \operatorname{int}(\operatorname{dom} f)$ and for all $x \in \operatorname{int}(\operatorname{dom} f)$,*

$$k(\nabla f(x_i)) \leq k(\nabla f(x)) - D_k(\nabla f(x), \nabla f(x_{i-1})) + L^* D_f(x_{i-1}, x) - L^* D_f(x_i, x). \tag{35}$$

Proof. Let $C = \operatorname{int}(\operatorname{dom} f)$. We proceed by induction. For $i = 0$ we have $x_0 \in C$ by assumption. Now, for $i > 0$, assume the induction hypothesis for x_{i-1}. Define,

$$x_\lambda = x_{i-1} - \frac{1}{\lambda} \nabla k(\nabla f(x_{i-1})) \tag{36}$$

for $\lambda > 0$. Because $x_{i-1} \in \operatorname{int}(\operatorname{dom} f) \neq \emptyset$, the set $S = \{\lambda \geq L^* : x_\lambda \in C\}$ is not empty. Let $x \in C$. Let $x^* = \nabla f(x)$, $x_{i-1}^* = \nabla f(x_{i-1})$ and $x_\lambda^* = \nabla f(x_\lambda)$ for $\lambda \in S$. The following identities follow by our definition of x_λ and some algebra.

$$\left\langle \nabla k(x_{i-1}^*), x^* - x_\lambda^* \right\rangle = \lambda \left\langle x_{i-1} - x_\lambda, x^* - x_\lambda^* \right\rangle, \tag{37}$$

$$\langle x_{i-1} - x_\lambda, \nabla f(x) - \nabla f(x_\lambda) \rangle = D_f(x_\lambda, x) + D_f(x_{i-1}, x_\lambda) - D_f(x_{i-1}, x). \tag{38}$$

Combining (37) and (38), we get

$$\begin{aligned}\lambda D_f(x_{i-1}, x_\lambda) + \left\langle \nabla k(x^*_{i-1}), x^*_\lambda - x^*_{i-1} \right\rangle = \\ \lambda D_f(x_{i-1}, x) - \lambda D_f(x_\lambda, x) + \left\langle \nabla k(x^*_{i-1}), x^* - x^*_{i-1} \right\rangle\end{aligned} \tag{39}$$

Putting everything together, we have

$$\begin{aligned}k(x^*_\lambda) &= k(x^*_{i-1}) + \left\langle \nabla k(x^*_{i-1}), x^*_\lambda - x^*_{i-1} \right\rangle + D_k(x^*_\lambda, x^*_{i-1}) \\ &\overset{(a)}{\leq} k(x^*_{i-1}) + \left\langle \nabla k(x^*_{i-1}), x^*_\lambda - x^*_{i-1} \right\rangle + L^* D_f(x_{i-1}, x_\lambda) \\ &\overset{(b)}{\leq} k(x^*_{i-1}) + \left\langle \nabla k(x^*_{i-1}), x^*_\lambda - x^*_{i-1} \right\rangle + \lambda D_f(x_{i-1}, x_\lambda) \\ &\overset{(c)}{=} k(x^*_{i-1}) + \left\langle \nabla k(x^*_{i-1}), x^* - x^*_{i-1} \right\rangle + \lambda D_f(x_{i-1}, x) - \lambda D_f(x_\lambda, x) \\ &\overset{(d)}{\leq} k(x^*) - D_k(x^*, x^*_{i-1}) + \lambda D_f(x_{i-1}, x) - \lambda D_f(x_\lambda, x).\end{aligned} \tag{40}$$

(a) follows from L^*-smoothness, (b) from $L^* \leq \lambda$ and the non-negativity of the Bregman divergence, (c) from (39), and (d) by definition and simple algebra. Taking $x = x_{i-1}$ and recalling the definition of x^*_{i-1} and x^*_λ reveals that

$$k(\nabla f(x_\lambda)) + \lambda D_f(x_\lambda, x_{i-1}) \leq k(\nabla f(x_{i-1})). \tag{41}$$

Now, our goal is to show that $x_i = x_{L^*} \in \operatorname{int}(\operatorname{dom} f)$ by showing that $L^* \in S$. We proceed by contradiction, so suppose $L^* \notin S$. Then $x_{L^*} \in \mathbb{R}^d \setminus \operatorname{int}(\operatorname{dom} f)$. Hence we can find $\Lambda \geq L^*$ such that $x_\Lambda \in \partial(\operatorname{dom} f)$. Now take a sequence $\lambda_j \to \Lambda$ such that $\lambda_j > \Lambda$. By the above discussion for all $j \geq 0$ we have $k(\nabla f(x_{\lambda_j})) \leq k(\nabla f(x_{i-1}))$. k being minimized at 0 means it satisfied Lemma 2.3 and thus is radially unbounded. This implies that $\left\|\nabla f(x_{\lambda_j})\right\| \leq c$ for some $c > 0$ and all $j \geq 0$. But this contradicts the requirement that $\left\|\nabla f(x_{\lambda_j})\right\| \to \infty$ since $x_{\lambda_j} \to x_\Lambda \in \partial(\operatorname{dom} f)$ from the assumption of essential smoothness. This completes the proof that $x_i = x_{L^*} \in \operatorname{int}(\operatorname{dom} f)$. Since $L^* \in S$, (40) ensures that (35) holds. □

We can now provide convergence rates for our method.

Theorem 3.9. *Let $f, k : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ satisfy Assumption 3.1. If $x_0 \in \operatorname{int}(\operatorname{dom} f)$, then for all $i > 0$ the iterates of Algorithm 1.1 satisfy*

$$k(\nabla f(x_i)) - k(0) \leq \frac{L^*}{i}(f(x_0) - f(x_{\min})). \tag{42}$$

In particular, $\nabla f(x_i) \to 0$. If additionally f is Legendre convex and there exists $\mu^* > 0$ such that $\mu^* D_f(x, y) \leq D_k(\nabla f(y), \nabla f(x))$ for all $x, y \in \operatorname{int}(\operatorname{dom} f)$, then for all $i > 0$ the iterates of Algorithm 1.1 satisfy

$$f(x_i) - f(x_{\min}) \leq \left(1 - \frac{\mu^*}{L^*}\right)^i (f(x_0) - f(x_{\min})). \tag{43}$$

Proof of Theorem 3.9. Let $C = \operatorname{int}(\operatorname{dom} f)$. We have $x_i \in C$ and $k(\nabla f(x_i)) \leq k(\nabla f(x_{i-1}))$ by the Descent Lemma 3.8. We also have $x_{\min} \in C$ by Lemma 2.2. Finally, (35) of Lemma 3.8 with $x = x_{\min}$ gives us,

$$\begin{aligned}k(\nabla f(x_i)) - k(0) &\leq L^*(f(x_{i-1}) - f(x_i)) - D_k(0, \nabla f(x_{i-1})) \\ &\leq L^*(f(x_{i-1}) - f(x_i))\end{aligned} \tag{44}$$

Putting this together, we get

$$\begin{aligned} i(k(\nabla f(x_i)) - k(0)) &\leq \sum_{j=1}^{i} k(\nabla f(x_i)) - k(0) \\ &\leq L^*(f(x_0) - f(x_i)). \end{aligned} \tag{45}$$

Dividing by i gives our first result. This implies that $k(\nabla f(x_i)) \to k(0)$, which implies that $\nabla f(x_i) \to 0$ by continuity and the uniqueness of k's minimum. Now, assume that f is Legendre convex and there exists $\mu^* > 0$ such that $\mu^* D_f(x, y) \leq D_k(\nabla f(y), \nabla f(x))$ for all $x, y \in \operatorname{int}(\operatorname{dom} f)$. For all $i > 0$,

$$\begin{aligned} L^*(f(x_i) - f(x_{\min})) &\overset{(a)}{\leq} L^*(f(x_{i-1}) - f(x_{\min})) - D_k(0, \nabla f(x_{i-1})) \\ &\overset{(b)}{\leq} L^*(f(x_{i-1}) - f(x_{\min})) - \mu^*(f(x_{i-1}) - f(x_{\min})), \end{aligned} \tag{46}$$

where (a) follows from (44) and the non-negativity of $k(x^*) - k(0)$. (b) follows from dual relative strong convexity. This inequality implies our desired result. □

Theorem 3.9 guarantees the convergence of the iterates of Algorithm 1.1 under the assumption that dual relative smoothness hold globally for a fixed L^*. Unfortunately it may be difficult to to derive a tight bound on L^* or small L^* may be appropriate locally. In this case, it may be useful to use a line search to choose L^*. Consider the following generalization of the update rule of Algorithm 1.1,

$$x_{i+1} = x_i - \frac{1}{L_i^*} \nabla k(\nabla f(x_i)) \tag{47}$$

where $L_i^* > 0$ is allowed to depend on the iteration. The next proposition shows that, under suitable assumptions, (47) converges with rates analogous to Theorem 3.9.

Proposition 3.10 (Adaptive step sizes)**.** *Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper closed convex function that is differentiable on $\operatorname{int}(\operatorname{dom} f) \neq \emptyset$ and minimized at $x_{\min}$. Let $k : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a proper closed convex function that is differentiable on $\nabla f(\operatorname{int}(\operatorname{dom} f))$. If $x_0 \in \operatorname{int}(\operatorname{dom} f)$ and for all $i > 0$ the iterates x_i in (47) satisfy*

1. *$x_i \in \operatorname{int}(\operatorname{dom} f)$,*
2. *$k(\nabla f(x_i)) \leq k(\nabla f(x_{i-1}))$,*
3. *$k(\nabla f(x_i)) - k(0) \leq L_{i-1}^*(f(x_{i-1}) - f(x_i))$,*

then we have

$$k(\nabla f(x_i)) - k(0) \leq \frac{\max_{0 \leq j \leq i-1} L_j^*}{i} (f(x_0) - f(x_{\min})). \tag{48}$$

Remark 3.11. In practice, a possible choice of step sizes is

$$L_{i-1}^* = \min\{2^r, r \in \mathbb{Z} : \text{ 1., 2., and 3. of Proposition 3.10 are satisfied}\}. \tag{49}$$

If L^* is the smallest real number such that f is dual L^*-smooth relative to k (see Lemma 3.6 for an equivalent condition), then this scheme satisfies that $L_{i-1}^* < 2L^*$ for every $i > 0$ (hence we are making steps that are almost as large or larger as if we would use the smallest possible fixed L^*, without knowing the value of L^* in advance). The search through the set in (49) for finding L_i^* can be initialized at L_{i-1}^*.

Proof of Proposition 3.10. The proof follows similar lines as in the previous case. First, by summing up the inequalities from 3, we obtain that

$$\sum_{1\leq j\leq i}[k(\nabla f(x_j))-k(0)]\leq \sum_{1\leq j\leq i} L_{i-1}^*(f(x_{i-1})-f(x_i))\leq (f(x_0)-f(x_{\min}))\max_{0\leq j\leq i-1} L_j^*,$$

and using 2., it follows that $\sum_{1\leq j\leq i}[k(\nabla f(x_j))-k(0)]\geq i(k(\nabla f(x_i))-k(0))$. The result follows directly. □

An important question that we do not address in this section is whether the sub-linear convergence of $k(\nabla f(x_i))-k(0)$ implies specific rates of convergence of other quantities of interest. These might be, for example, $\|x_i-x_{\min}\|$ or $f(x_i)-f(x_{\min})$. Rates for these will likely depend on both f and k.

4 Applications

4.1 Exponential Penalty Functions

Consider the following problem.

$$\min_{x\in\mathbb{R}^d}\{c^Tx : Ax\leq b\}, \tag{LP}$$

where $c\in\mathbb{R}^d$, $b\in\mathbb{R}^n$, and $A\in\mathbb{R}^{n\times d}$. Associate with this linear program the following relaxation into an unconstrained problem: $\min_{x\in\mathbb{R}^d} f_\tau(x)$ for

$$f_\tau(x)=c^Tx+\tau\sum_{i=1}^n \exp((A_ix-b_i)/\tau), \tag{50}$$

where $\tau>0$ and A_i is the ith row of A (a row vector). This approximation of (LP) with exponential penalty functions was studied by several authors (see [44, 21, 37, 5]) and is directly useful in the machine learning literature for boosting (see, e.g., [32]). Here we design a dual reference function for f_τ under the following assumptions.

Assumption 4.1. *Suppose that the following hold for problem* (LP).

1. $\|A_i\|=1$ *for* $1\leq i\leq n$.
2. $A\in\mathbb{R}^{n\times d}$ *is of full rank* $d\leq n$.
3. $P=\{x\in R^n : Ax\leq b\}$ *is a polytope, which is contained in a Euclidean ball of radius* $R>0$ *and contains a Euclidean ball of radius* $r>0$.

The dual reference function will be designed so that is smooth relative to f_τ^* and Algorithm 1.1, with appropriate step-size choices, converges with global guarantees.

Define the dual reference function $k:\mathbb{R}^d\to\mathbb{R}$,

$$k(x^*)=\|x^*\|-\log(\|x^*\|+1). \tag{51}$$

This behaves like a quadratic $\|x^*\|^2/2$ near its minimum $x^* = 0$ and like $\|x^*\|$, i.e., grows linearly, at infinity. It is also possible to verify that k is Legendre convex. Furthermore, we have:

$$\nabla k(x^*) = \frac{x^*}{\|x^*\| + 1}, \qquad \nabla^2 k(x^*) = \frac{I}{\|x^*\| + 1} - \frac{x^* x^{*T}}{(\|x^*\| + 1)^2 \|x^*\|}. \tag{52}$$

Hence, $[\nabla^2 k(x^*)]^{-1} \succeq (1 + \|x^*\|)I$. From Proposition 3.4 and this inequality it follows that the fact that k is L^*-smooth relative to f_τ^* is implied by

$$\nabla^2 f_\tau(x) \preceq L^* \left[1 + \|\nabla f_\tau(x)\|\right] I \qquad \forall x \in \mathbb{R}^d. \tag{53}$$

This is the strategy of the following theorem, which shows that f_τ is dual smooth to this choice of k under our assumptions.

Proposition 4.2. *Under Assumption 4.1 for f_τ defined in* (50) *and k defined in* (51)*, we have that*

$$\nabla^2 f_\tau(x) \preceq L_\tau^* [\nabla^2 k(\nabla f_\tau(x))]^{-1} \qquad \forall x \in \mathbb{R}^d, \tag{54}$$

where the dual relative smoothness constant is given by

$$L_\tau^* = \frac{2R}{r} \frac{\left\|A^T A\right\|}{\tau} (\eta + \|c\|). \tag{55}$$

Here, $\left\|A^T A\right\|$ is the induced matrix norm, and

$$\eta = \sup_{\|s\|_\infty \leq 1} \left\|A^T s\right\| \leq \sqrt{n} \left\|A^T\right\|_\infty. \tag{56}$$

Because f_τ and k are Legendre convex, k is smooth relative to f_τ^* and Theorem 3.9 implies that Algorithm 1.1 converges with $k(\nabla f(x_i))$ converging at a rate $\mathcal{O}(1/i)$.

Remark 4.3. From Theorem 3.9, we have

$$k(\nabla f_\tau(x)) \leq \frac{L_\tau^* (f_\tau(x_0) - f_\tau(x_{\min}))}{i}. \tag{57}$$

This suggests that, if we can start from an initial point within the polytope, then we can reach a point where $\|\nabla f_\tau(x)\|$ is significantly less than $\|c\|$ (which is expected to be near the minimum) in polynomial amount of steps, depending on the conditioning R/r and the value of τ. The step-size $1/L_i^*$ can also be chosen adaptively, as explained in Proposition 3.10. Near the minimum, both $f_\tau(x)$ and $k(x^*)$ behave like quadratic functions, so local linear convergence rates hold. We believe that this iterative scheme is reasonably efficient for high dimensional well-conditioned polytopes, but in other less well conditioned instances it is outperformed by existing algorithms such as multiplicative weights [4] or [20], which is based on Newton's method (hence uses second-order information).

Proof of Proposition 4.2. Note that $1 \leq \eta \leq n$, because $\|A_i\| = 1$. Let $\alpha(x) := \max_{i \in [n]} (A_i x - b_i)$. Then $\alpha(x) < 0$ inside the polytope and $\alpha(x) > 0$ outside of it. By differentiation, we have

$$\nabla f_\tau(x) = \sum_{i=1}^n A_i \exp((A_i x - b_i)/\tau) + c, \tag{58}$$

$$\nabla^2 f_\tau(x) = \sum_{i=1}^n \frac{A_i^T A_i}{\tau} \exp((A_i x - b_i)/\tau). \tag{59}$$

Note that f_τ is defined everywhere and differentiable. Furthermore, under our assumption that $\operatorname{rank}(A^TA) = \operatorname{rank}(A) = d$, it is evidently strictly convex and therefore Legendre.

The Hessian of f_τ satisfies

$$\nabla^2 f_\tau(x) \preceq \exp(\alpha(x)/\tau)\frac{A^TA}{\tau} \preceq \exp\left(\alpha(x)/\tau\right)\frac{\left\|A^TA\right\|}{\tau} I. \tag{60}$$

Because $\eta \geq 1$, it is clear that the claim of the theorem holds for every x where $\alpha(x) \leq 0$ (i.e. inside the polytope or on its boundary). From now on we will assume that x is such that $\alpha(x) > 0$ (outside of the polytope). Let x_c be a minimizer of $\alpha(x)$ (at least one exists since the polytope is compact and $\alpha(x)$ is a continuous function), then using the assumption $\|A_i\| = 1$ it follows that $\alpha(x_c) = -r < 0$. Hence $x \neq x_c$. We are going to need an upper bound on $\|x - x_c\|$, which we will obtain as follows. By the definitions, we have $A_i x_c \leq -r + b_i$ and $A_i x = A_i x - b_i + b_i \leq \alpha(x) + b_i$, hence

$$\begin{aligned} A_i\left(x_c + \frac{r}{\alpha(x)+r}(x - x_c)\right) &= \frac{r}{\alpha(x)+r}A_i x + \frac{\alpha(x)}{\alpha(x)+r}A_i x_c \\ &\leq \frac{r}{\alpha(x)+r}(\alpha(x) + b_i) + \frac{\alpha(x)}{\alpha(x)+r}(-r + b_i) = b_i. \end{aligned}$$

Therefore $x_c + \frac{r}{\alpha(x)+r}(x - x_c) \in P \subset B_{x_c}(2R)$, so

$$0 < \|x - x_c\| \leq 2\frac{\alpha(x)+r}{r}R \quad \text{and} \quad \|x - x_c\|^{-1} \geq \frac{r}{\alpha(x)+r}\frac{1}{2R}. \tag{61}$$

Let $\mathcal{I} = \{i \in [n];\, A_i x - b_i > 0\}$, $\mathcal{J} = \{i \in [n];\, A_i x - b_i \leq 0\}$, and

$$G_{\mathcal{I}}(x) = \sum_{i\in\mathcal{I}} e^{\frac{1}{r}(A_i x - b_i)} A_i \qquad G_{\mathcal{J}}(x) = \sum_{i\in\mathcal{J}} e^{\frac{1}{r}(A_i x - b_i)} A_i. \tag{62}$$

Then $\nabla f_\tau(x) = G_{\mathcal{I}}(x) + G_{\mathcal{J}}(x) + c$. We have

$$\begin{aligned} \|G_{\mathcal{I}}(x)\| \geq \frac{G_{\mathcal{I}}(x)^T(x - x_c)}{\|x - x_c\|} &= \|x - x_c\|^{-1} \sum_{i\in\mathcal{I}} e^{\frac{1}{r}(A_i x - b_i)} A_i(x - x_c) \\ &\overset{(a)}{\geq} \|x - x_c\|^{-1} e^{\frac{\alpha(x)}{\tau}}(\alpha(x) + r) \\ &\overset{(b)}{\geq} \frac{r}{2R} e^{\frac{\alpha(x)}{\tau}}. \end{aligned}$$

Here, (a) follows from the facts that there is a $j \in \mathcal{I}$ such that $A_j(x - x_c) = \alpha(x) + b_j - A_j x_c \geq \alpha(x) + r$ and the fact that $A_i(x - x_c) \geq b_i + r - b_i > 0$ holds for every $i \in \mathcal{I}$. (b) follows from (61). From (60) we obtain that

$$\begin{aligned} \nabla^2 f_\tau(x) &\preceq \exp\left(\alpha(x)/\tau\right)\frac{\left\|A^TA\right\|}{\tau} I \\ &\preceq \frac{2R}{r}\frac{\left\|A^TA\right\|}{\tau}\|G_{\mathcal{I}}(x)\|\, I \\ &\preceq \frac{2R}{r}\frac{\left\|A^TA\right\|}{\tau}(\|\nabla f_\tau(x)\| + \|G_{\mathcal{J}}(x)\| + \|c\|) I. \end{aligned} \tag{63}$$

Hence (53) follows from the facts that $\|G_{\mathcal{J}}(x)\| \leq \eta$ and $\eta + \|c\| \geq 1$. As discussed (54) follows from $[\nabla^2 k(x^*)]^{-1} \succeq (1 + \|x^*\|) I$. □

4.2 p-norm Regression

Consider the following p-norm regression problem,

$$\min_{x\in\mathbb{R}^d} \|Ax-b\|_p^p, \tag{pnorm}$$

where $A\in\mathbb{R}^{n\times d}$, $d\ll n$, $b\in\mathbb{R}^n$, and $p\geq 1$. This problem is a useful abstraction for some important graph problems, including Lipschitz learning on graphs [29] and ℓ_p-norm minimizing flows [1]. Algorithms specialized for p-norm regression have recently been studied in the theoretical computer science literature by several authors (see, e.g., [17, 2] and references therein). In this subsection, we design an appropriate dual reference function for (pnorm) under the following assumptions. Let A_i denote the rows of A (as row vectors).

Assumption 4.4. *Suppose that the following hold for problem* (pnorm).

1. $2\leq p<\infty$.
2. *A is full rank d, and for all $x\in\mathbb{R}^d$ there is a subset $I(x)\subset[n]$ such that $A_ix\neq b_i$ for all $i\in I(x)$, and* $\operatorname{span}\{A_i : i\in I(x)\}=\mathbb{R}^d$.
3. $c_G=\inf_{\|s\|=1}\|As\|_p^p>0$.
4. $c_H=\inf_{u,v\in\mathbb{R}^d:\|u\|=1,\|v\|=1}\sum_{i=1}^n |A_iu|^{p-2}(A_iv)^2>0$.

Remark 4.5. Although these assumptions seem restrictive, we can show that, if $n\geq 2d-1$ and $(A_i)_{1\leq i\leq n}$ and $(b_i)_{1\leq i\leq n}$ are chosen as independent random variables with densities that are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and $\mathbb{R}$, then the assumptions hold with probability 1. Assumption 2 is implied by the stronger assumption that any d rows of A define a full rank d matrix, and the maximal number of equalities $A_ix=b_i$ that hold for any x is no more than d. This stronger version of Assumption 2, and Assumption 3 holds with probability 1 under the random allocation due to the fact that the set of real valued $d\times d$ matrices with determinant 0 has Lebesgue-measure 0 in $\mathbb{R}^{d\times d}$ (due to the fact that the determinant is a multivariate polynomial of the entries, and the zero set of such polynomials has Lebesgue measure zero unless they are constant 0, see [19]). The minimum in Assumption 4 is achieved for some $u_{\min}$ and $v_{\min}$ due to continuity and compactness of the unit sphere. Since any d rows of A form an independent basis with probability 1, it follows that u and v can be orthogonal to at most $d-1$ of them, respectively, so using $n\geq 2d-1$ there exists an i in the sum $\sum_{i=1}^n |A_iu_{\min}|^{p-2}(A_iv_{\min})^2$ that is non-zero, hence Assumption 4 holds.

Consider the dual reference function $k:\mathbb{R}^d\to\mathbb{R}$,

$$k(x^*)=\tfrac{1}{q}\left(\|x^*\|^2+1\right)^{\frac{q}{2}}-\tfrac{1}{q}, \tag{64}$$

for $q=\frac{p}{p-1}$ (hence $\frac{1}{p}+\frac{1}{q}=1$). This behaves like a quadratic $\|x^*\|^2/2$ near its minimum $x^*=0$ and like $\|x^*\|^q/q$ at infinity. For this k, we have

$$\nabla k(x^*)=x^*(1+\|x^*\|^2)^{\frac{q-2}{2}} \tag{65}$$

As the next theorem shows dual relative strong convexity and smoothness of k relative to the conjugate of (pnorm) hold under our assumptions.

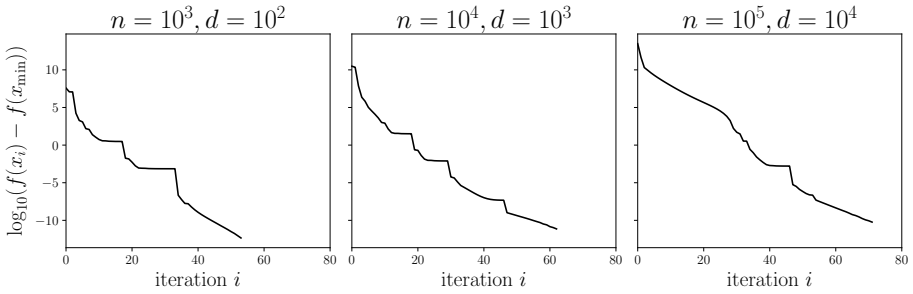

Figure 1: Convergence rates for p-norm regression are mostly unaffected by the dimension d for these random instances with $p = 4$.

Proposition 4.6. *Let $f(x) = \|Ax - b\|_p^p$ be the p-norm objective. Under Assumption 4.4 for k defined in* (64)*, there exists $\mu^*, L^* > 0$ such that*

$$\mu^*[\nabla^2 k(\nabla f(x))]^{-1} \preceq \nabla^2 f(x) \preceq L^*[\nabla^2 k(\nabla f(x))]^{-1} \qquad \forall x \in \mathbb{R}^d. \tag{66}$$

See (76) *and* (77) *for the definitions of μ^* and L^*. Because f and k are Legendre convex, k is smooth and strongly convex relative to f^* and Theorem 3.9 implies that Algorithm 1.1 converges with $f(x_i) - f(x_{\min})$ converging at a linear rate $\mathcal{O}((1 - \mu^*/L^*)^i)$.*

To test the empirical performance of this method, we have implemented it with A_i, b, and x_0 i.i.d. as standard normals for power $p = 4$, $d \in \{10^2, 10^3, 10^4\}$, and $n = 10d$. The inverse step-size L_0^* was chosen to be $L_0^* = 1$ initially, and multiplied by 2 if the function value would increase due to too large steps (hence this was chosen adaptively in the beginning, but L_i^* was never decreased later on). As Figure 1 shows, empirically our method seems to be performing well, with high precision achieved after 50-80 gradient evaluations, and the convergence rate seems to be mostly unaffected by the dimension d. Hence in this random setting dual space preconditioning is indeed very efficient, and competitive with previous works [17, 2, 1] which had dimension dependent convergence rates. We think that based on Proposition 4.6, it can be shown that with high probability, dimension-free convergence rates hold in this random scenario when the number of vectors n tends to infinity (the proof would be based on concentration inequalities for empirical processes, see e.g. [14] for an overview of such inequalities). Note however that we do not believe this always to be the case for general non-random A and b, and there could be instances of very poor conditioning (such as when $n \approx d$) where the homotopy method of [17] or the IRLS method of [3] could perform better. The proof of Proposition 4.6 is based on the following two lemmas.

Lemma 4.7 (Bounds on the gradient)**.** *Let $f(x) = \|Ax - b\|_p^p$ be the p-norm objective for* (pnorm)*. Under Assumption 4.4, we have*

$$L_G\|x\|^{p-1} - C_G \leq \|\nabla f(x)\| \leq U_G\|x\|^{p-1} + D_G \tag{67}$$

for all $x \in \mathbb{R}^d$, with constants

$$L_G = 2^{-p+1} c_G = 2^{-p+1} \inf_{\|s\|=1} \|As\|_p^p, \quad C_G = \left(\sum_{i=1}^n |b_i|^p \right)^{(p-1)/p} \cdot c_G^{1/p},$$
$$U_G = 2^{p-2}(p+1) \sup_{\|s\|=1} \|As\|_p^p, \quad D_G = 2^{p-2}(p-1) \left(\sum_{i=1}^n |b_i|^p \right).$$

Proof. By differentiation, we have

$$\nabla f(x) = p \sum_{i=1}^n |A_i x - b_i|^{p-2} (A_i x - b_i) A_i, \tag{68}$$

thus

$$\begin{aligned}
\|\nabla f(x)\| &= p \left\| \sum_{i=1}^n |A_i x - b_i|^{p-2} (A_i x - b_i) A_i \right\| \\
&\geq \max \left(\frac{p}{\|x\|} \sum_{i=1}^n |A_i x - b_i|^{p-2} (A_i x - b_i) A_i x, 0 \right) \\
&= \max \left(\frac{p}{\|x\|} \sum_{i=1}^n \left[|A_i x - b_i|^{p-2} (A_i x - b_i)^2 + |A_i x - b_i|^{p-2} (A_i x - b_i) b_i \right], 0 \right) \\
&\geq \max \left(\frac{p}{\|x\|} \sum_{i=1}^n \left(|A_i x - b_i|^p - |A_i x - b_i|^{p-1} |b_i| \right), 0 \right),
\end{aligned}$$

now by Young's inequality $|A_i x - b_i|^{p-1} |b_i| \leq |A_i x - b_i|^p \frac{p-1}{p} + \frac{|b_i|^p}{p}$, hence

$$\geq \max \left(\frac{1}{\|x\|} \sum_{i=1}^n \left(|A_i x - b_i|^p - |b_i|^p \right), 0 \right)$$

using the fact that $|a+b|^p \leq (|a| + |b|)^p = \left(\frac{2|a| + 2|b|}{2} \right)^p \leq 2^{p-1}(|a|^p + |b|^p)$ by convexity (this is so-called the C_p inequality), so $|A_i x - b_i|^p + |b_i|^p \geq 2^{-p+1} |A_i x|^p$, hence

$$\begin{aligned}
&\geq \max \left(\frac{1}{\|x\|} \sum_{i=1}^n \left(2^{-p+1} |A_i x|^p - 2|b_i|^p \right), 0 \right) \\
&\geq \max \left(2^{-p+1} \left[\inf_{\|s\|=1} \|As\|_p^p \right] \cdot \|x\|^{p-1} - \frac{2 \sum_{i=1}^n |b_i|^p}{\|x\|}, 0 \right),
\end{aligned}$$

and the lower bound follows from Assumption 4.4 by straightforward rearrangement. For the upper

bound, notice that

$$\begin{aligned}
\|\nabla f(x)\| &\leq p \sup_{\|v\|=1} \sum_{i=1}^{n} |A_i x - b_i|^{p-1} |A_i v| \\
&\leq 2^{p-2} p \sup_{\|v\|=1} \sum_{i=1}^{n} \left(|A_i x|^{p-1} |A_i v| + |b_i|^{p-1} |A_i v| \right) \\
&\leq 2^{p-2} p \left[\|x\|^{p-1} \sup_{\|s\|=1, \|v\|=1} \sum_{i=1}^{n} \left(|A_i s|^{p-1} |A_i v| \right) + \sup_{\|v\|=1} \sum_{i=1}^{n} |b_i|^{p-1} |A_i v| \right] \\
&\leq 2^{p-2} p \left[\frac{p+1}{p} \sup_{\|s\|=1} \|As\|_p^p + \frac{p-1}{p} \sum_{i=1}^{n} |b_i|^p \right],
\end{aligned}$$

hence the result follows. The last step uses Fenchel-Young, and rearrangement. □

Lemma 4.8 (Bounds on the Hessian)**.** *Let $f(x) = \|Ax - b\|_p^p$ be the p-norm objective. Suppose that Assumption 4.4 holds, and let*

$$R_H = \left\| \sum_{i=1}^{n} |b_i|^{p-2} A_i^T A_i \right\|^{1/(p-2)} / (c_H 2^{-p})^{1/(p-2)}, \tag{69}$$

$$\rho_H = \inf_{\|x\| \leq R_H} \lambda_{\min}(\nabla^2 f(x)) = \inf_{\|x\| \leq 1, \|u\|=1} p(p-1) \sum_{i=1}^{n} |A_i x - b_i|^{p-2} (A_i u)^2. \tag{70}$$

Then $\rho_H > 0$, and we have

$$(L_H \|x\|^{p-2} + C_H) I \preceq \nabla^2 f(x) \preceq (U_H \|x\|^{p-2} + D_H) I \tag{71}$$

for all $x \in \mathbb{R}^d$, with constants

$$\begin{aligned}
L_H &= \min \left(p(p-1) 2^{-p-1} c_H, \frac{\rho_H}{2 R_H^{p-2}} \right), \\
U_H &= 2^{p-3} p(p-1) \sup_{\|u\|=1, \|v\|=1} \sum_{i=1}^{n} |A_i u|^{p-2} (A_i v)^2, \\
C_H &= \min \left(\frac{\rho_H}{2}, p(p-1) 2^{-p-1} c_H R_H^{p-2} \right), D_H = p(p-1) 2^{p-3} \left\| \sum_{i=1}^{n} |b_i|^{p-2} A_i^T A_i \right\|.
\end{aligned}$$

Proof. We have by differentiation

$$\nabla^2 f(x) = p(p-1) \sum_{i=1}^{n} |A_i x - b_i|^{p-2} A_i^T A_i. \tag{72}$$

Notice that using the fact that $|a-b|^{p-2}+|b|^{p-2} \geq 2^{-(p-1)}|a|^{p-2}$, we have

$$\begin{aligned}\nabla^2 f(x) &= p(p-1)\sum_{i=1}^{n} |A_i x - b_i|^{p-2} A_i^T A_i \\ &\succeq p(p-1)\sum_{i=1}^{n} \left(2^{-(p-1)}|A_i x|^{p-2} - |b_i|^{p-2}\right) A_i^T A_i \\ &\succeq p(p-1)2^{-(p-1)}c_H \|x\|^{p-2} - p(p-1)\left\|\sum_{i=1}^{n} |b_i|^{p-2} A_i^T A_i\right\|.\end{aligned}$$

Let R_H be as in (69), then using the above bound, we can see that for $\|x\| \geq R_H$, we have

$$\begin{aligned}\nabla^2 f(x) &\succeq p(p-1)2^{-p}c_H \|x\|^{p-2} I \\ &\succeq p(p-1)2^{-p-1}c_H \|x\|^{p-2} + p(p-1)2^{-p-1}c_H R_H^{p-2}.\end{aligned} \tag{73}$$

Since the minimum of the continuous function $\lambda_{\min}(\nabla^2 f(x))$ is achieved on the compact set B_{R_H}, and by the second part of Assumption 4.4, it cannot be zero, and hence $\rho_H > 0$ and $\nabla^2 f(x) \succeq \rho_H I$ for every $x \in B_{R_H}$. The lower bound in (71) follows by combining this with (73). For the upper bound, using the inequality $|a+b|^{p-2} \leq 2^{p-3}(|a|^{p-2}+|b|^{p-2})$, we obtain that

$$\begin{aligned}\nabla^2 f(x) \preceq\ & p(p-1)2^{p-3} \sup_{\|s\|=1} \left\|\sum_{i=1}^{n} |A_i s|^{p-2} A_i^T A_i\right\| \cdot \|x\|^{p-2} \\ &+ p(p-1)2^{p-3}\left\|\sum_{i=1}^{n} |b_i|^{p-2} A_i^T A_i\right\|.\end{aligned}$$

□

Now we are ready to prove our main result in this section.

Proof of Proposition 4.6. First, both f and k are Legendre convex in this case. This is easy to verify for k, and evidently f is differentiable everywhere. To verify strict convexity of f, note that $\nabla^2 f(x) \succ 0$ under part two of Assumption 4.4. Since both f and k are twice differentiable, by Proposition 3.4, it suffices to check that (66) holds for the linear convergence of Algorithm 1.1. We have by differentiation,

$$\nabla^2 k(x^*) = (1+\|x^*\|^2)^{\frac{q-2}{2}} I + (q-2)(1+\|x^*\|^2)^{\frac{q-4}{2}} x^* x^{*T}. \tag{74}$$

Now it is easy to see that for $p \in [2,\infty)$, we have $q = p/(p-1) \in (1,2]$ and it is not difficult to verify that $\nabla^2 k$ satisfies that for all $x^* \in \mathbb{R}^d$,

$$(1+\|x^*\|^2)^{\frac{1}{2}\frac{p-2}{p-1}} I \preceq \left[\nabla^2 k(x^*)\right]^{-1} \preceq (p-1)(1+\|x^*\|^2)^{\frac{1}{2}\frac{p-2}{p-1}} I. \tag{75}$$

The claim of the theorem now follows by some straightforward rearrangement using Lemmas 4.7 and 4.8, with constants

$$\mu^* = \min\left(\frac{C_H}{2(p-1)(2+2D_G)}, \frac{L_H}{4(p-1)U_G^{(p-2)/(p-1)}}\right), \tag{76}$$

$$L^* = \min\left(\frac{U_H}{(L_G/2)^{(p-2)/(p-1)}}, 4U_H\left(\frac{C_G}{L_G}\right)^{(p-2)/(p-1)} + 2D_H\right). \tag{77}$$

□

5 Discussion

In this paper we introduced a non-linear preconditioning scheme for gradient descent on Legendre convex functions f that converges under generalizations of the standard Lipschitz assumption on ∇f. There are at least two interpretations of this method. The first is as a generalization of gradient descent in which the update direction is preconditioned by the gradient map ∇k of a designed dual reference, Legendre convex function k. The second interpretation is as a Bregman gradient method in the dual space, which minimizes the designed k while the conjugate f^* plays the role of the "reference function". The choice of k affects the conditioning of our method, which is made explicit in our analysis through a relative smoothness condition between k and f^*. The dual relative conditions admit non-smooth f and k, and are provably distinct dual cousins of the relative smoothness conditions introduced by [7]. k serves as a model of the convex conjugates f^* in a certain problem class. In section 4, we show how this method can be applied to exponential penalty functions (see, e.g., [21, 20]) and p-norm regression (see [17, 2] and references therein) with global convergence rate guarantees.

Algorithm 1.1 is related to a number of existing methods, some of which are subject to the analysis we provide. The most notable of these is the method of steepest descent with respect to a given norm $\|\cdot\|$ (now not necessarily Euclidean). Here we follow the exposition of Boyd and Vandenberghe [15, sect. 4.9]. The steepest descent iteration is given by

$$x_{i+1} = x_i + \frac{1}{L} \|\nabla f(x_i)\|_* d, \qquad \text{where } d \in \arg\max_{\|x\| \leq 1} \langle -\nabla f(x_i), x \rangle , \tag{78}$$

and $\|x^*\|_* = \sup_{\|x\| \leq 1} \langle x, x^* \rangle$ is the dual norm of $\|\cdot\|$. The identity $\partial(\|x^*\|_*^2/2) = \|x^*\|_* \arg\max\{\langle x^*, x\rangle : \|x\| \leq 1\}$ for all $x^* \in \mathbb{R}^d$ implies that for strictly convex and differentiable $\|\cdot\|_*$, the steepest descent method (78) is a special case of dual preconditioned gradient descent with $k(x^*) = \|x^*\|_*^2/2$. Our analysis does not apply in the case of other norms or normalized steepest descent [15]. Algorithm 1.1 also generalizes the rescaled gradient method of [47, sect. 2.2]. Thus, our method may be seen as a generalization of the steepest descent method and rescalings of gradient descent. Dual preconditioning is more distantly related to the dual gradient methods [43, 9]. These methods are designed for problems with non-smooth, but strongly convex structure. They exploit the duality between classical smoothness and strong convexity by applying smooth minimization algorithms to a dual problem. Similarly, Algorithm 1.1 can be seen as a move to the dual space, in which a dual problem $k(x^*) \approx f^*(x^*) - \langle x^*, x_{\min} \rangle$ (dual to $f(x) + \delta_{x=x_{\min}}(x)$) is minimized by a Bregman gradient method. Thus, dual gradient methods and dual preconditioning are most easily applied when the dual structure is relatively more benign to model than the primal structure, e.g., when f has super-quadratic growth.

There are a couple natural questions that arise from this work. First, it may be useful to pursue the analogy with dual gradient methods further and to design methods for the general composite model that exploit dual relative smoothness. Second, there is still considerable difficulty in the design of k. Thus, it may be productive to investigate whether methods from linear preconditioning (see [10] for a review), such as incomplete factorizations or sparse approximate inverses, can be generalized to the non-linear setting for the design of k. Nonetheless, the dual relative conditions

studied in this work provide new avenues for improving the conditioning of optimizers via hard-won domain-specific knowledge.

Acknowledgements

We thank the anonymous referees for their insightful comments that helped us to improve the paper. We thank David Balduzzi for insightful comments, Patrick Rebeschini for suggesting exponential penalty functions and Sushant Sachdeva for suggesting p-norm regression. CJM acknowledges the support of the Institute for Advanced Study and the James D. Wolfensohn Fund, the support of a DeepMind Graduate Scholarship, and the support of the Natural Sciences and Engineering Research Council of Canada under reference number PGSD3-460176-2014. YWT's research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) ERC grant agreement no. 617071. This material is based upon work supported in part by the U.S. Army Research Laboratory and the U. S. Army Research Office, and by the U.K. Ministry of Defence (MoD) and the U.K. Engineering and Physical Research Council (EPSRC) under grant number EP/R013616/1.

References

[1] Deeksha Adil and Sushant Sachdeva. Faster p-norm minimizing flows, via smoothed q-norm problems. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 892–910. SIAM, 2020.

[2] Deeksha Adil, Rasmus Kyng, Richard Peng, and Sushant Sachdeva. Iterative refinement for ℓ_p-norm regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1405–1424. SIAM, 2019.

[3] Deeksha Adil, Richard Peng, and Sushant Sachdeva. Fast, provably convergent irls algorithm for p-norm linear regression. In *Advances in Neural Information Processing Systems*, pages 14189–14200, 2019.

[4] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory Comput.*, 8(1):121–164, 2012.

[5] Alfred Auslender, Roberto Cominetti, and Mounir Haddou. Asymptotic analysis for penalty and barrier methods in convex and linear programming. *Math. Oper. Res.*, 22(1):43–62, 1997.

[6] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *J. Mach. Learn. Res.*, 6:1705–1749, 2005.

[7] Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res.*, 42(2): 330–348, 2016.

[8] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, 2003.

[9] Amir Beck and Marc Teboulle. A fast dual proximal gradient algorithm for convex minimization and applications. *Oper. Res. Lett.*, 42(1):1–6, 2014.

[10] Michele Benzi. Preconditioning techniques for large linear systems: a survey. *J. Comput. Phys.*, 182(2):418–477, 2002.

[11] Benjamin Birnbaum, Nikhil R. Devanur, and Lin Xiao. New convex programs and distributed algorithms for fisher markets with linear and spending constraint utilities. Technical Report MSR-TR-2010-112, August 2010.

[12] Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM J. Optim.*, 28(3):2131–2151, 2018.

[13] Jonathan Borwein and Adrian S Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer Science & Business Media, 2010.

[14] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[15] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[16] Lev Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

[17] Sébastien Bubeck, Michael B Cohen, Yin Tat Lee, and Yuanzhi Li. An homotopy method for lp regression provably beyond self-concordance and in input-sparsity time. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1130–1137. ACM, 2018.

[18] Xiao-Chuan Cai and David E Keyes. Nonlinearly preconditioned inexact newton algorithms. *SIAM J. Sci. Comput.*, 24(1):183–200, 2002.

[19] Richard Caron and Tim Traynor. The zero set of a polynomial. Technical report, 2005. Available at `http://www1.uwindsor.ca/math/sites/uwindsor.ca.math/files/05-03.pdf` and `https://www.researchgate.net/publication/281285245_The_Zero_Set_of_a_Polynomial`.

[20] Roberto Cominetti and Jean-Pierre Dussault. Stable exponential-penalty algorithm with superlinear convergence. *J. Optim. Theory Appl.*, 83(2):285–309, 1994.

[21] Roberto Cominetti and Jaime San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Math. Program.*, 67(1-3):169–187, 1994.

[22] Victorita Dolean, Martin J Gander, Walid Kheriji, Felix Kwok, and Roland Masson. Nonlinear preconditioning: How to use a nonlinear schwarz method to precondition newton's method. *SIAM J. Sci. Comput.*, 38(6):A3357–A3380, 2016.

[23] Radu-Alexandru Dragomir, Jérôme Bolte, and Alexandre d'Aspremont. Fast gradient methods for symmetric nonnegative matrix factorization. *arXiv e-prints*, January 2019.

[24] Nicolas Flammarion and Francis Bach. Stochastic composite least-squares regression with convergence rate $o(1/n)$. In *Conference on Learning Theory*, 2017.

[25] Anne Greenbaum. *Iterative methods for solving linear systems*. SIAM, 1997.

[26] Magnus R Hestenes et al. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952.

[27] Martin Hutzenthaler, Arnulf Jentzen, Peter E Kloeden, et al. Strong convergence of an explicit numerical method for sdes with nonglobally lipschitz continuous coefficients. *Ann. Appl. Probab.*, 22(4):1611–1641, 2012.

[28] Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic.uchicago.edu/~shai/papers/KakadeShalevTewari09.pdf*, 2009.

[29] Rasmus Kyng, Anup Rao, Sushant Sachdeva, and Daniel A Spielman. Algorithms for lipschitz learning on graphs. In *Conference on Learning Theory*, pages 1190–1223, 2015.

[30] Haihao Lu. "Relative Continuity" for Non-Lipschitz Nonsmooth Convex Optimization Using Stochastic (or Deterministic) Mirror Descent. *INFORMS J. Optim.*, 2019.

[31] Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.*, 28(1):333–354, 2018.

[32] Ron Meir and Gunnar Rätsch. An introduction to boosting and leveraging. In *Advanced lectures on machine learning*, pages 118–183. Springer, 2003.

[33] Konstantin Mishchenko. Sinkhorn algorithm as a special case of stochastic mirror descent. *arXiv e-prints*, Sep 2019.

[34] Arkadi Nemirovski and David Yudin. Effective methods for the solution of convex programming problems of large dimensions. *Ekonom. i Mat. Metody*, 15(1):135–152, 1979.

[35] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[36] Yurii Nesterov. Implementable tensor methods in unconstrained convex optimization. *Math. Program.*, pages 1–27, 2019.

[37] Roman Polyak and Marc Teboulle. Nonlinear rescaling and proximal-like methods in convex optimization. *Math. Program.*, 76(2):265–284, 1997.

[38] R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[39] Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

[40] Sotirios Sabanis. A note on tamed Euler approximations. *Electron. Commun. Probab.*, 18, 2013.

[41] Shai Shalev-Shwartz and Yoram Singer. Online learning: Theory, algorithms, and applications. 2007.

[42] Marc Teboulle. A simplified view of first order methods for optimization. *Math. Program.*, 170 (1):67–96, 2018.

[43] Paul Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM J. Control Optim.*, 29(1):119–138, 1991.

[44] Paul Tseng and Dimitri P Bertsekas. On the convergence of the exponential multiplier method for convex programming. *Math. Program.*, 60(1-3):1–19, 1993.

[45] Quang Van Nguyen. Forward-backward splitting with bregman distances. *Vietnam J. Math.*, 45(3):519–539, 2017.

[46] A. J. Wathen. Preconditioning. *Acta Numer.*, 24:329–376, 2015. doi: 10.1017/S0962492915000021.

[47] Ashia Wilson, Lester Mackey, and Andre Wibisono. Accelerating rescaled gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

[48] Constantin Zălinescu. *Convex Analysis in General Vector Spaces*. World Scientific, 2002.